Kin Selection and Its Critics

Jonathan Birch and Samir Okasha

October 6, 2014

Abstract: Hamilton's theory of kin selection is the best-known framework for understanding the evolution of social behaviour, but has long been a source of controversy in evolutionary biology. A recent critique of the theory by Nowak, Tarnita and Wilson sparked a new round of debate, which shows no signs of abating. In this overview we highlight a number of conceptual issues that lie at the heart of the current debate. We begin by emphasizing that there are various alternative formulations of Hamilton's rule, including a 'general' version that is always true, an 'approximate' version that assumes weak selection, and a 'special' version that demands other restrictive assumptions. We then examine the relationship between the 'neighbour-modulated fitness' and 'inclusive fitness' approaches to kin selection. Finally, we consider the often strained relationship between the theories of kin and multi-level selection.

Keywords: Hamilton's rule, social evolution, kin selection, inclusive fitness, multi-level selection

---

Introduction

The pithiest expression of the concept of kin selection was made long before the theory itself was devised, when J.B.S. Haldane is said to have quipped "I would lay down my life for two brothers or eight cousins". The remark captures an intuitive and powerful thought: when interacting organisms share genes, they may have an evolutionary incentive to help each other. Moreover, and more profoundly, it suggests that the size of the incentive to help is proportional to the degree of relatedness between them. We owe the formal embodiment of this insight to Hamilton (1964), and the term 'kin selection' to Maynard Smith (1964). Today, Hamilton's theory lies

at the heart of an established and sizeable research program, the explanatory domain of which has steadily expanded (Bourke 2011a).

The basic empirical prediction of kin selection theory is that social behaviour should correlate with genetic relatedness; in particular, 'altruistic' actions – which are costly to the actor but benefit others – are more likely to be directed towards relatives. This qualitative prediction has been amply confirmed in diverse taxa, including microbes, insects and vertebrates. Moreover, kin selection has shed light on a range of biological phenomena including dispersal, sex-ratio allocation, worker-queen conflicts in insect colonies, the distribution of reproduction in animal societies ('reproductive skew'), parasite virulence, genomic imprinting, the evolution of multicellularity, and more (Bourke 2011a). The principles of kin selection also help illuminate aspects of the 'major transitions in evolution', which occur when free-living individuals coalesce to form a new higher-level entity which eventually becomes an 'individual' itself (Maynard Smith and Szathmáry 1995; Bourke 2011a).

Despite its empirical success, kin selection theory is not without its critics. For example E. O. Wilson, the famous author of *Sociobiology*, was once an enthusiastic supporter of kin selection but has changed his mind. In their recent work on eusocial insect colonies, Wilson and his co-author Bert Hölldobler argue that genetic relatedness is less important than is often thought; on their view, ecological factors, rather than high levels of within-colony relatedness, are the primary drivers of the evolution of eusociality (Wilson and Hölldobler 2005, Hölldobler and Wilson 2008).

In August 2010, a strongly-worded critique of kin selection by Nowak, Tarnita and Wilson (2010) ignited a new round of debate in *Nature*. In March 2011, a rebuttal was published signed by 137 social evolution theorists, who claimed that Nowak and colleagues' arguments "are based on a misunderstanding of evolutionary theory, and a misrepresentation of the empirical literature" (Abbot et al. 2011, p. E1). More detailed rebuttals have since appeared (Rousset and Lion 2011; Gardner et al. 2011; Bourke 2011b), plus a response by Nowak and colleagues (Nowak et al. 2011). Follow-up critiques by van Veelen et al. (2012), Wilson (2012), Allen et al. (2013) and Wilson and Nowak (2014) have left continuing uncertainty about the status of Hamilton's theory. Does it lie in tatters? Or is it alive and kicking, healthier than ever? It depends on who you ask.

In this overview, we offer a fresh look at some of the issues raised by this debate. As philosophers of science, rather than practising biologists, we hope to bring a certain detachment

to the discussion. Our aim is not to debunk or vindicate kin selection, nor to take a stand on any empirical questions, but to offer some conceptual clarifications. In Section 2, we discuss the core explanatory principle of kin selection theory, 'Hamilton's rule'. We emphasize that although the name suggests a single, unambiguous principle, there are in fact various formulations of the rule which it is crucial to distinguish. In Section 3, we examine the relationship between the 'neighbour-modulated fitness' and 'inclusive fitness' approaches to kin selection, and look briefly at the idea that inclusive fitness is the quantity that organisms should appear designed to maximize. In Section 4, we examine the often strained relationship between the theories of kin and group selection, and ask whether these theories are ultimately equivalent, as is often claimed. In Section 5, we close by highlighting some outstanding issues.

The status of Hamilton's rule

The central explanatory principle of kin selection theory is 'Hamilton's rule', which says that a gene coding for a social behaviour will be favoured by natural selection if and only if $rb > c$, where b represents the 'benefit' the behaviour confers on the recipient, c represents the 'cost' it imposes on the actor, and r is the 'coefficient of relatedness' between actor and recipient (Hamilton 1964). The costs and benefits are measured in increments of reproductive fitness. The rule tells us that an altruistic behaviour will be favoured by selection so long as the fitness cost to the actor is offset by a sufficient amount of benefit to sufficient closely related recipients.

In contemporary discussions, r is intended to encompass any relevant genetic similarity between actors and recipients, regardless of the mechanism that led to it. Hence although Hamilton originally defined r in genealogical terms, as a measure of shared ancestry, in principle Hamilton's rule still applies when genetic correlations arise by other means, including 'greenbeard' effects (Dawkins 1976; Gardner and West 2010), pleiotropic effects (Hamilton 1975), and, in microbes, gene mobility (Mc Ginty et al. 2013; Birch 2014a). In practice, however, genealogical kinship remains the most common source of genetic correlation between social partners.

In their 2010 paper, Nowak et al. (2010) say that Hamilton's rule "almost never holds" (p. 1059), in the sense that it almost never constitutes a true statement of the conditions under which a social behaviour will be favoured by natural selection. This claim elicited vigorous rebuttals from their opponents—most notably from Gardner, West and Wild (2011), who retort that "it is simply incorrect to claim that Hamilton's rule requires restrictive assumptions or that

it almost never holds" (p. 1038). There is, at present, no sign of an end to this divisive dispute (see Nowak et al. 2011; Allen et al. 2013; West and Gardner 2013). It is hard to see how both camps can be right, yet neither seems likely to budge.

Three versions: HRS, HRG and HRA

The key to understanding the current standoff is to see that, when social evolution theorists talk about 'Hamilton's rule', they may have a number of subtly different principles in mind. Hamilton (1964) first derived a result of the form '$rb > c$' in a one-locus population-genetic model that made a number of substantial assumptions, including weak selection, additive gene action (i.e. no dominance or epistasis) and the additivity of fitness payoffs (i.e. a relatively simple payoff structure). In the following decades, numerous theorists (including Hamilton himself) explored the extent to which a similar result could be recovered when some or all of Hamilton's original assumptions were relaxed. The upshot was a variety of different routes to '$rb > c$'-type results, often with contrasting implications about the conditions under which the rule applies (e.g. Hamilton 1975; Michod 1982; Grafen 1985; Queller 1984, 1992; Frank 1998, 2013; Rousset 2004; Lehmann and Keller 2006; Lehmann and Rousset 2010, 2014a, b).

Within this rather bewildering space of alternative formulations of Hamilton's rule, one three-way distinction is particularly salient. It concerns the meaning of the 'cost' and 'benefit' coefficients. First of all, there are formulations in which 'cost' and 'benefit' denote *payoff parameters* of a specific evolutionary model. Examples include the formulations of Queller (1984), Taylor and Nowak (2007), van Veelen (2009), Nowak et al. (2010) and van Veelen et al (2012). Second, there are formulations in which the 'cost' and 'benefit' terms are partial regression coefficients (i.e. 'average effects', in the sense of Fisher 1941) which quantify the *overall statistical associations* in a population between an organism's genotype/phenotype, its fitness, and the genotype/phenotype of social partners—which can in principle be computed for any model or set of population data. Queller's (1992) formulation is one example, recently defended and applied by Gardner et al. (2007, 2011). Third, there are formulations in which 'cost' and 'benefit' refer to marginal, first-order *approximations* of regression coefficients. This is the approach most commonly used by contemporary kin selection theorists. Roughly speaking (since this is not the place for detailed mathematical exposition), the approximation works by replacing differences with differentials. That is, it approximates the regression coefficients corresponding to c and b with partial derivatives of a fitness function (Taylor and Frank 1996;

Frank 1998, 2013; Rousset 2004; Lehmann and Rousset 2010, 2014a, b).

Some clear labels will help us keep these versions apart. For the exact version of the rule in which c and b are payoff parameters, we suggest the name 'HRS' ('S' for special). For the exact, regression-based version of Queller (1992), we suggest the name 'HRG' ('G' for general). For the marginal approximation of HRG, we suggest the name 'HRA' ('A' for 'approximate').

Which version we have in mind makes an important difference to the generality of Hamilton's rule. HRS is an exact result for any model with an additive payoff structure—that is, a payoff structure in which the payoff an actor's behaviour confers on a recipient is independent of the recipient's phenotype and combines with other payoffs by adding up. This, however, amounts to a significant restriction. It is easy to construct counterexamples to HRS simply by considering a non-additive payoff structure in which the payoff a given social action confers on a recipient *does* depend on the recipient's own phenotype. This point was noted by Queller (1984) and has recently been emphasized by van Veelen (2009). Unsurprisingly, when the payoff structure of social interaction is too complex to represent with just two parameters (as is the case in non-additive scenarios), a rule more complicated than HRS is needed to describe the condition for a social behaviour to spread (Queller 1984; van Veelen 2009).

However, if we define 'c' and 'b' as partial regression coefficients (as in HRG), we obtain a version of Hamilton's rule of much greater generality. Indeed, we end up with an exact version of the rule that remains correct no matter how complicated the payoff structure may be, because all relevant payoff parameters are implicitly taken into account in the calculation of cost and benefit (Queller 1992; Gardner et al. 2007, 2011). In effect, this is because we are abstracting away from the complex causal details of social interaction to focus on the overarching statistical relationship between genotype and fitness. This generalized, regression-based version of Hamilton's rule is always true because it makes no assumptions at all about how these statistical relationships are mediated phenotypically.

The marginal approximation of HRG (i.e. HRA) sacrifices a degree of this generality, since the approximation of differences by differentials is justified only if selection is weak and gene action is additive (Frank 1998; Lehmann and Rousset 2014b). However, HRA does not presuppose an additive payoff structure, and it thus holds (unlike HRS) across a wide range of game-theoretic scenarios. The key is to note that HRA is fundamentally an approximate result. Rather than assuming that the payoff structure is additive, HRA relies on the idea that, when selection is weak, a first-order approximation that neglects deviations from payoff addi-

tivity is justified. In broad terms, then, HRA provides an intermediate degree of generality. Its assumptions are more restrictive than those of HRG, but less restrictive than those of HRS.

We can use this three-way distinction to make sense of the ongoing standoff. When Nowak et al. (2010) say that "Hamilton's rule almost never holds", they are referring to HRS, the exact version of the rule in which c and b refer to payoff parameters. Meanwhile, when Gardner et al. (2011) say that "it is simply incorrect to say that Hamilton's rule requires restrictive assumptions or almost never holds", they are referring to the exact, regression-based version employed by Queller (1992), Gardner et al. (2007) and others. Once we distinguish HRS from HRG, we see that both of these apparently contradictory statements are correct (Birch 2014b). Neither statement here is referring to HRA, even though this approximate version of the rule is the version most commonly used by kin selection theorists.

Does HRG explain anything?

Getting clear about the definitions of 'cost' and 'benefit' does not wholly resolve the conflict over Hamilton's rule, because underneath the terminological fog of war there are substantive issues at stake. One question is whether, if (as in HRG) we define the c and b terms so that Hamilton's rule is always true, we buy generality at the cost of explanatory power. As far as Nowak et al. are concerned, HRG adds nothing to our understanding of social evolution:

> There are attempts to make Hamilton's rule work by choosing generalized cost and benefit parameters [HRG], but these parameters are no longer properties of individual phenotypes. They depend on the entire system including population structure. These extended versions of Hamilton's rule have no explanatory power for theory or experiment. (Nowak et al. 2011)

Do Nowak et al. have a case? It is undoubtedly true that HRG has predictive limitations. For example, one might expect Hamilton's rule to predict that if we were to intervene to increase the genetic relatedness between social partners, cooperative behaviour would be more likely to evolve. But there are simple models in which the r, c, and b coefficients in HRG are all interdependent, with the result that intervening to increase relatedness also increases the cost/benefit ratio, making cooperative behaviour less likely to evolve. Similarly, one might intuitively predict that if a social behaviour satisfies Hamilton's rule at one time, it will continue to do so in the future, provided there is no change in the underlying payoff structure or the

relatedness between social partners. But the c and b coefficients in HRG will typically depend on population gene frequency—with the consequence that a social behaviour may satisfy HRG at a low frequency but not at a higher frequency (Allen et al. 2013; Birch 2014b; Lehmann and Rousset 2014a).

These concerns about the predictive limitations of HRG are real, but do not imply that it has no explanatory power at all. This is because, although prediction and explanation are related, they are not exactly the same thing. As philosophers of science have often noted, a principle can be explanatory without being predictive, and vice versa (Salmon 1989). In the philosophy of science, there is a long tradition of pointing to unification as an important aspect of scientific explanation (Kitcher 1989). In this spirit, some defenders of HRG have argued that it constitutes a unifying principle in social evolution theory which helps us see what otherwise disparate models have in common (Gardner et al. 2007; Birch 2014b).

However, in addition to its unifying power, Hamilton's rule is often also taken to embody an important *causal* insight about social evolution, namely that a costly social behaviour will spread only if the direct fitness *effect* of the behaviour on the actor who performs it is outweighed by the indirect fitness *effect* on the recipient, weighted by the relatedness between them, where 'effect' is understood causally and not just statistically. This causal interpretation of HRG is valid only if the 'c' and 'b' regression coefficients admit of an interpretation as causal effects. It is not entirely clear when it is legitimate to interpret them in this way, because there is no general theory of when exactly a partial regression coefficient (or Fisherian 'average effect') admits of a causal interpretation. The debate is on-going, and connects in interesting ways to debates surrounding Fisher's fundamental theorem (Lee and Chow 2013). What we do know, however, is that, partial regression coefficients are certainly not causally interpretable in *all* cases (Spirtes et al. 2000; Queller 2011; Allen et al. 2013; Birch 2014b). To think otherwise is to confuse causation and correlation. Indeed, Allen et al. (2013) provide several hypothetical examples in which a causal interpretation of the coefficients is not reasonable.

By this point, it is clear that the debate has taken on a partly philosophical character, turning on subtle issues concerning the relation between causality and statistics, and the explanatory function that Hamilton's rule is intended to serve. These are issues that neither mathematical modelling nor empirical studies can decisively settle. For this reason, debates about the value of HRG are unlikely to go away. But if researchers manage to steer clear of semantic confusions fostered by the alternative formulations of Hamilton's rule, then there is room for a constructive

debate regarding the rule's explanatory uses and limitations.

The status of inclusive fitness

Hamilton's original 1964 paper introduced the concept of 'inclusive fitness', a modification of the classical fitness concept for dealing with social interactions. An organism's inclusive fitness is defined as a weighted sum, over all individuals in the population (including itself), of those portions of each individual's reproductive output for which the organism is causally responsible, with the weights given by relatedness coefficients. Hamilton observed that an altruistic action, which by definition will reduce an organism's personal fitness, may nonetheless enhance its inclusive fitness; and he proposed that social evolution be understood as a process of inclusive fitness maximization. The status of the inclusive fitness concept is another bone of contention in the current controversy. Nowak et al. (2010) and Allen et al. (2013) argue that the concept has no advantages over the traditional fitness concept. By contrast, Grafen (2006), Bourke (2011a) and West and Gardner (2013) argue that inclusive fitness is the key to understanding social evolution.

Neighbour-modulated and inclusive fitness

Inclusive fitness is not the only way to formulate kin selection theory. As Hamilton himself noted, an alternative is to use 'neighbour-modulated fitness', which is in some ways a more intuitive notion. To see the difference between them, consider two viewpoints on what happens when altruism evolves by virtue of relatedness between social partners (Box 1). One is to view relatedness as a source of *correlated interaction*: when r is high, bearers of the genes for altruism are differentially likely to interact with other bearers, hence to receive the benefits of other agents' altruism. Thus high r means that bearers of the genes for altruism may have greater reproductive success, on average, than non-bearers. The other is to view relatedness as a source of *indirect reproduction*: when r is high, recipients provide actors with an indirect means of securing genetic representation in the next generation. Thus genes for altruism may spread, if the indirect representation an altruist secures through helping its relatives exceeds the representation it loses through sacrificing a portion of its own reproduction success.

The first perspective is captured in the 'neighbour-modulated fitness' framework (Figure 1), which looks at the correlations between an individual's genotype and its social neighbourhood,

and helps predict when these correlations will make bearers of the genes for altruism fitter, on average, than non-bearers (Hamilton 1964; Taylor and Frank 1996; Frank 1998, 2013). The second perspective is captured in the 'inclusive fitness' framework (Figure 2), which adds up all the fitness effects causally attributable to a social actor—weighting each component by the relatedness between actor and recipient—in order to calculate the net effect of a social behaviour on the actor's overall genetic representation in the next generation (Hamilton 1964; Frank 1998, 2013; Grafen 2006).

[BOX 1 GOES HERE]

[FIGURE 1 GOES HERE]

[FIGURE 2 GOES HERE]

Although correlated interaction and indirect reproduction may sound like different mechanisms, the inclusive and neighbour-modulated fitness frameworks are usually considered equivalent, as they generally yield identical results about when a social behaviour will evolve (Taylor et al. 2007). Thus the choice is one of modelling convenience, not empirical fact. Hamilton (1964) and Maynard Smith (1983) both regarded inclusive fitness as easier to apply in practice. But in recent years this situation has largely reversed: kin selection theorists have increasingly come to favour the neighbour-modulated fitness framework, citing *its* greater simplicity and ease of application (Taylor and Frank 1996; Taylor et al. 2007; Gardner et al. 2007).

In one respect, the neighbour-modulated approach is more general. To perform an inclusive fitness analysis, we need to be able to attribute each social phenotype to a single controlling genotype (Frank 1998). By contrast, a neighbour-modulated fitness analysis simply ignores the pathway from actor genotypes to social phenotypes, leaving us with one fewer causal path to worry about. A corollary is that the neighbour-modulated framework can apply in cases where there is no principled way to ascribe a social character to a single controlling genotype. As Frank (1998, 2013) notes, cases in which phenotypes are controlled by actors of a different species to the recipient—such as host-parasite interaction—arguably fall into this category (though cf. Taylor et al. 2007).

Inclusive fitness and the 'objective' of social behaviour

One advantage of the inclusive fitness approach is that it helps to make precise the idea that organisms' social behaviour is 'purposive', or goal-oriented. This idea of purpose, or apparent purpose, is a key component of the 'adaptationist' approach to evolution that Darwin initiated. Where non-social traits are concerned, biologists typically assume that an evolved trait will serve to enhance an organism's expected reproductive output; models based on the assumption often enjoy empirical success. But altruistic behaviours seemingly do not fit this paradigm, as they reduce rather than enhance an organism's personal fitness. It is here that the inclusive fitness concept comes into its own, allowing us to rescue the idea that social behaviour should appear purposive by suitably re-defining the 'purpose' in question, namely enhancement of inclusive, rather than personal, fitness. This feature of the inclusive fitness concept explains its popularity among behavioural ecologists, and has been emphasized in recent work by Grafen (2006, 2014), Gardner, West and Wild (2011), Okasha, Weymark and Bossert (2014) and others.

What enables inclusive fitness to play this role is its focus on which actors control which phenotypes. Recall that an actor's inclusive fitness is a relatedness-weighted sum of the fitness effects for which it is causally responsible. Thus we can put ourselves in the position of the actor and ask: 'How should I behave, in order to maximize my expected inclusive fitness?' Since natural selection tends to favour traits that promote inclusive fitness on average, this question can serve as an informal route to predictions of which social behaviours will evolve. By contrast, we cannot usefully ask the same question with regard to neighbour-modulated fitness, because an individual's neighbour-modulated fitness contains components over which it may have no control. All we can do is put ourselves in the position of a recipient and ask: 'What genotypes are "good news", as far as my neighbour-modulated fitness is concerned?' But this heuristic is considerably less intuitive, because considerations of causation and control are replaced by considerations of statistical auspiciousness.

The idea that social behaviour should serve to maximize an organism's inclusive fitness is hinted at in Hamilton's original 1964 papers but not made fully explicit. In his recent work on the 'formal Darwinism project', Alan Grafen has attempted to place the idea on a firm footing, by proving formal links between gene-frequency change and an 'optimization program' (Grafen 2006, 2014). Essentially, Grafen seeks to prove, in a quite general setting, that if all the organisms in a population choose an action (from a fixed set of possible actions) that maximizes their inclusive fitness, then population-genetic equilibrium will obtain; and vice-versa. While

(as Grafen admits) this falls short of proving that natural selection will always lead inclusive fitness maximizing behaviour to evolve (e.g. because gene frequencies may cycle indefinitely), it arguably provides some support for that belief. In effect, Grafen's results (taken at face value) mean that so long as the population does actually evolve towards a stable equilibrium, then we should expect inclusive-fitness maximizing behaviour to evolve.

Grafen's results rest on one key assumption, namely that costs and benefits have additive phenotypic effects on fitness. This means, for example, that the benefit b that an altruistic action has on the recipient is independent of the recipient's own genotype. In general this is not a realistic assumption, as it rules out any frequency-dependence of fitness, though it may be a good approximation in certain cases. Whether Grafen's results can be extended to the non-additive case has not yet been settled; see Lehmann and Rousset (2014a) and Gardner, West and Wild (2011) for conflicting opinions on this issue.

At this point it is useful to recall the 'general' formulation of Hamilton's rule (HRG), which as we saw defines the r, b and c co-efficients in such a way that the $rb > c$ condition is always correct, irrespective of whether costs and benefits are additive or not. It is tempting to suggest that Grafen's optimization results could be extended to the non-additive case, and thus made fully general, simply by defining inclusive fitness using the r, b and c terms of HRG. However there is a problem with this suggestion. For recall that an organism's inclusive fitness is supposed to be fully within its control, i.e. to depend only the social actions that it performs. Since the b and c terms of HRG are functions of population-wide gene frequencies, the amount of inclusive fitness an organism gets from a given action would depend on the state of the population, if inclusive fitness were defined as suggested.

This suggests that the generalization of Grafen's results on inclusive fitness maximization to the non-additive case will be difficult to achieve. Further, it highlights the important difference between Hamilton's rule itself—the statement of the conditions under which an allele for a social behaviour will be favoured by selection—and the idea that an organism's evolved behaviour will serve to maximize its inclusive fitness. These two aspects of kin selection theory, though related, should be kept distinct.

Kin selection and multi-level selection

Another dimension of the current controversy concerns the relation between kin and multi-level (or group) selection. Kin and multi-level selection provide seemingly quite different perspectives on social evolution. Kin selection, as we have seen, emphasizes the relatedness between social partners as the crucial factor mediating the spread of a pro-social behaviour. Multi-level selection, by contrasts, emphasizes the interplay of selection within groups and between groups (Price 1972; Hamilton 1975; Sober and Wilson 1998; Okasha 2006). Within any group, altruists will be at a selective disadvantage vis-à-vis their selfish counterparts; but groups containing a high proportion of altruists may outcompete groups containing a lower proportion. So for an altruistic behaviour to spread, the between-group component of selection must trump the within-group component.

The relation between kin and multi-level selection has been a source of controversy ever since it was first broached by Hamilton (1975). In earlier debates biologists tended to regard kin and multi-level selection as rival empirical hypotheses (e.g. Maynard Smith 1964, 1976; Dawkins 1976). But many contemporary biologists regard them as ultimately equivalent, on the grounds that gene frequency change can be correctly computed using either (e.g. Marshall 2011; Lehmann et al. 2007; Frank 2013). Although dissenters from this equivalence claim can be found (e.g. Hölldobler and Wilson 2009; van Veelen 2009; Traulsen 2010; Nowak et al. 2010), the majority of social evolutionists appear to endorse it.

Formal equivalence

To see the grounds for the equivalence claim, consider a simple model. A population of haploid individuals live in groups of the same size, within which social interactions occur (Figure 3). An allele at a particular locus codes for a social behaviour. Define $p_i = 1$ if the i^{th} individual has the allele, and $p_i = 0$ otherwise. The index i ranges over all individuals in the global population, irrespective of group membership. The population-wide frequency of the allele is $\overline{p}$. The reproductive output ('fitness') of individual i, defined as the total number of surviving offspring it contributes to the next generation, is denoted w_i. The average fitness in the population is $\overline{w}$. Mutation is assumed absent.

[FIGURE 3 GOES HERE]

Under these assumptions, the change in allele frequency over a single generation is given by:

$$\overline{w}\Delta\overline{p} = \mathrm{Cov}(w_i, p_i) \tag{1}$$

This is a version of the Price equation (Price 1970); the full version includes an extra term, but we are entitled to drop that term here because our assumptions guarantee the unbiased transmission of alleles. The equation tells us that the allele, and thus the social behaviour that it codes for, will spread so long as $\mathrm{Cov}(w_i, p_i) > 0$, i.e. there is a positive covariance between an individual's fitness and its genetic value. This simply formalizes the core neo-Darwinian idea that genes associated with higher individual fitness will increase in frequency.

Equation (1) is always true but not always useful, as the covariance term will often lack a natural biological interpretation (Grafen 2006; Okasha forthcoming). Kin and multi-level selection can be regarded as alternative ways of decomposing the covariance term in (1) into more meaningful components. On the kin selection approach, we use a linear regression model to split the covariance term into components attributable to the direct and indirect fitness effects of the social behaviour under consideration (Queller 1992; Gardner et al. 2011). This allows us to straightforwardly derive HRG, the generalized version of Hamilton's rule discussed above. On the multi-level selection approach, we split the covariance term into components attributable to selection within groups and selection between groups (Price 1972; Okasha 2006). This allows us to derive a principle that closely parallels HRG, according to which a costly social behaviour can spread by natural selection only if the selection for the trait between groups is stronger than the selection against the trait within groups. The details of these derivations are spelled out in Boxes 2 and 3.

[BOX 2 GOES HERE]

[BOX 3 GOES HERE]

We can now see why kin and multi-level selection are often regarded as equivalent. In any group-structured population, the total evolutionary change can be decomposed using either the kin selection partition (equation 4) or the multi-level partition (equation 6). Moreover, it is easy to see that the kin selection criterion for spread of a pro-social trait ($rb > c$), will be satisfied

if and only if the multi-level criterion (between-group > within-group) is satisfied. Thus the two approaches are *formally* equivalent. Gene frequency change can be computed in two ways: by determining the magnitude of the between and within-group components, or the direct and indirect effects; both methods will always give the same answer. In effect, the two approaches can be seen as alternative ways of caputing the fundamental insight that positive assortment, i.e. altruists interacting preferentially with each other, is what is crucially needed for altruism to evolve.

Recently, van Veelen (2009) and van Veelen et al. (2012) have challenged the received wisdom on this issue, arguing that the kin and multi-level selection are *not* formally equivalent, and that the latter is in fact more general than the former; see also Traulsen (2010). The HRS/HRG distinction introduced above again helps us understand what is going on here. What van Veelen et al. have shown, in effect, is that the *special* version of Hamilton's rule, HRS, is not formally equivalent to the standard multi-level decomposition in Box 3. This is true but should come as no surprise, since HRS applies only under restrictive assumptions. Their argument does not threaten the equivalence results of Marshall (2011) and others, because these results concern the formal equivalence of the multi-level selection approach and the *general* version of Hamilton's rule. Again, the key is to distinguish between the maximally general formulation of kin selection, i.e. HRG, and more specific formulations.

In one respect, the kin selection approach is arguably more general than the multi-level approach. For the latter requires that individuals are nested into non-overlapping groups, as in Figure 3 above; this is necessary for the decomposition technique in Box 3 to apply (Hamilton 1975; Okasha 2006; Frank 2013). Groups of this sort exist in some taxa, e.g. the colonies of many social insect species. But in other cases, individuals engage in social interactions with their conspecifics but there are no well-defined, discrete groups. The kin selection approach can handle such cases easily; indicative of this is that in deriving equation (4) above (Box 2), we did not make use of the fact that the individuals were nested into non-overlapping groups. Thus the claim that kin and multi-level selection are 'formally equivalent' requires at least this qualification.

Choosing between them

On a practical, day-to-day basis, social evolution researchers must decide which approach to use—and the formal equivalence of the two approaches does not imply that there is no principled

basis on which to choose between them. West et al. (2008) are emphatic on this point:

> At one level, kin selection and group selection are just different ways of doing the maths or conceptualizing the evolutionary process. However, from a practical point of view, it could not be clearer that the kin selection approach is the more broadly applicable tool that we can use to understand the natural world. This is because kin selection methodologies are usually easier to use, allow the construction of models that can be better linked to specific biological examples, lend themselves to empirical testing and allow the construction of a general conceptual overview. In addition, the group selection approach is not only less useful, but also appears to frequently have negative consequences by fostering confusion that leads to wasted effort (West et al. 2008, pp. 381-382).

Is this a fair assessment? It is true that the kin selection approach (in both its neighbour-modulated and inclusive fitness guises) has received more theoretical attention than the group selection approach, and has been put to work in more empirical applications. For example, kin selection models can straightforwardly take into account class structure, whereby different types of social agent in a population have different reproductive value (Taylor 1990; Frank 1998); and they are readily hooked up with the Taylor-Frank method, a powerful technique for the prediction of evolutionarily stable strategies (Taylor and Frank 1996; Frank 1998). However this does not show that the multi-level approach is unworthy of a similar degree of theoretical attention or that it is inherently unsuited to empirical applications. Indeed, given that kin and multi-level selection are formally rather similar—they simply partition up the total evolutionary change in slightly different ways—claims that one approach is inherently superior to the latter, as proponents of each have argued, must be treated with a degree of scepticism.

The widespread preference for kin selection may be partly due to multi-level selection's association with the flawed 'good of the group' tradition of the 1950s and 1960s, and the associated 'superorganism' concept of which many biologists remain suspicious. It is undeniable that the careless appeal to group-level advantage as a way of explaining a trait's evolution led to serious errors in the past; so biologists' wariness of this mode of explanation is understandable. Kin selection is an 'individualistic' methodology that makes no explicit mention of group fitness or group advantage, so has often seemed preferable for that reason (e.g. Dawkins 1976). However, this consideration should not be overplayed. Past errors not notwithstanding, multi-level

selection has evolved into a respectable theory, and does not necessarily carry a commitment to the superorganism concept (which is, at best, defensible only in special cases such as clonal groups or highly advanced eusocial insect colonies; cf. Gardner and Grafen 2009; Okasha and Paternotte 2012). Moreover, the idea that kin selection is methodologically preferable to multi-level selection seems hard to square with their formal equivalence. Indeed, those who have been favoured kin selection on these grounds have typically not properly appreciated that equivalence (West et al. 2008 is an exception in this respect).

It has recently been suggested that kin selection has a unique advantage over multi-level selection in that it comes with an associated 'optimization principle' (Gardner and Grafen 2009; Gardner et al. 2011). The suggestion here is that the concept of organisms maximizing their inclusive fitness, which permits social behaviour to be brought within the Darwinian paradigm, is the key insight of kin selection theory, but has no good parallel in multi-level selection theory. The putative parallel would presumably involve groups maximizing their 'group fitness', but this notion only makes sense for fully clonal groups, it has been argued (Gardner and Grafen 2009, though cf. Okasha and Patternote 2012). This line of argument is interesting but not conclusive, given that the circumstances in which it has been shown that evolution will lead individuals to maximize their inclusive fitness are anyway fairly restricted, as emphasized above.

Causal aptness

Finally, we want to suggest a different sort of consideration that might help biologists choose between the kin and multi-level approaches in a given context. The basic thought is that, although kin and multi-level selection are equivalent as *statistical* decompositions of evolutionary change, there are situations in which one approach provides a more accurate representation of the *causal structure* of social interaction. For evolutionary biology, like other sciences, is interested in constructing causal explanations; ideally we want our descriptions of evolutionary change to capture the causal structure of the underlying selection process, as well as correctly computing allele frequency change. So although kin and multi-level selection may be formally equivalent, it does not follow that they are also equally good as causal representations.

For example, suppose we are investigating a segregation distorter allele which also has deleterious effects on the fitness of its bearer. It is very natural to describe the selection pressures operating on this allele in multi-level terms: at the gene level, there is selection in its favour; but at the organism level, there is selection against it. The formal equivalence of kin and group

selection suggests that, if we wanted, we could re-describe the whole situation in terms of the inclusive fitness interests of the allele, but it is not clear what we stand to gain in explanatory terms by doing so. On the contrary, this move would seem unhelpful: it would obscure the true causal structure of the scenario, which clearly involves two distinct levels of selection. When we are looking at selection occuring both between and within organisms, a multi-level description seems clearly more apt, causally speaking.

However, there are other cases in which a kin selection description seems more apt from a causal point of view. Consider a Prisoner's Dilemma-style scenario in which organisms interact in pairs and must choose whether to cooperate or defect. Suppose that genetic correlation between social partners leads to the evolution of cooperation. It seems natural to describe this in terms of kin selection: to say, for example, that organisms cooperate because it is in their inclusive fitness interests to do so. As Sober and Wilson (1998) point out, however, any such scenario may be re-described in the language of multi-level selection. For if we regard each interacting pair as a group of size 2, we can say that within each group defectors outperform cooperators, but groups with more cooperators outperform groups with fewer. Yet as in the previous example, it is not clear what we stand to gain from this rather strained description of the process. After all, these 'groups of size 2' may be highly ephemeral, coming into existence when the social interaction begins and vanishing as soon as it is complete. If this is the case, then they are 'groups' in name only, and describing this as a process of multi-level selection seems to sow confusion rather than insight.

Plainly, our intuitions about these two examples do not constitute a full-blown theory of causal aptness; they do not provide any general recipe for deciding which description is causally superior in any given case. Nevertheless, they are enough to show that considerations of causal aptness do matter, if we want our theories and models of social evolution to embody causal—as opposed to merely statistical—truths. Developing a more adequate treatment of causal aptness remains an important direction for future work. Okasha (forthcoming) attempts a systematic analysis of the circumstances under which kin and multi-level selection offer better causal representations of social evolution, using tools from the theory of causal modelling (Pearl 2009).

Conclusions and open questions

There are many oustanding issues in the foundations of social evolution theory. We feel that progress on these issues is achieveable if rival camps of researchers are able to communicate and cooperate, rather than pursuing divergent research programs. In this overview we have tried to take an even-handed approach that identifies what both critics and defenders of kin selection have got right, while highlighting the ways in which theorists have at times talked past one another. We will close by highlighting three questions that we hope future work in this area will address.

Q1: When do the c and b coefficients in HRG admit of a causal interpretation?

In Section 2, we noted that the generalized version of Hamilton's rule, HRG, defines the c and b coefficients using the statistical concept of regression. In effect, in applying HRG, we are fitting a plane to a three-dimensional cloud of population data describing each organism's genotype, its social partner's genotype, and its fitness; c and b are the coefficients which specify that plane. But can HRG tell us anything about the causal processes involved in the evolution of social behaviour, given that it is defined in purely statistical terms? For as Allen et al. (2013) have pointed out, following Spirtes et al. (2000), there are many cases in which regression coefficients should *not* be interpreted causally. The issue lies at the heart of the ongoing debates surrounding Hamilton's rule, but a systematic treatment is currently lacking.

Q2: How widely applicable is the idea that evolution will lead individuals to 'try' to maximize their inclusive fitness?

In Section 3, we noted that inclusive fitness appears to offer an 'objective' for social behaviour, as it is a quantity that is within the 'control' of the individual actor. However, the most careful attempt to justify the idea that evolution in social contexts will lead individuals to behave as if trying to maximize their inclusive fitness, due to Grafen (2006), rests on assumptions that severely limit its generality. It is currently unclear whether Grafen's argument, or one like it, can be extended to cover non-additive scenarios and to cover frequency-dependent selection.

Q3: Under what conditions are kin and multi-level selection causally, as opposed to formally, equivalent?

In Section 4, we noted that kin and multi-level selection, when formulated in general terms as alternative decompositions of the Price equation, are formally equivalent in that allele frequency change can be correctly computed in both ways. But intuitively, there are cases in which one is more causally apt than the other. However, a general account of causal aptness that goes beyond our intuitions in simple cases has yet to be constructed.

Acknowledgements

SO was supported by the European Research Council Seventh Framework Program (FP7/20072013), ERC Grant agreement no. 295449.

References

Abbot P, and 136 others. 2011. Inclusive fitness theory and eusociality. Nature 471: E1-E2.

Allen B, Nowak MA, Wilson EO. 2013. Limitations of inclusive fitness. Proceedings of the National Academy of Sciences USA. Forthcoming. doi: 10.1073/pnas.1317588110

Birch J. 2014a. Gene mobility and the concept of relatedness. Biology and Philosophy 29: 445-476.

Birch J. 2014b. Hamilton's rule and its discontents. British Journal for the Philosophy of Science 65: 381-411.

Bourke AFG. 2011a. Principles of Social Evolution. Oxford University Press.

Bourke AFG. 2011b. The validity and value of inclusive fitness theory. Proceedings of the Royal Society of London B: Biological Sciences 278: 3313-3320.

Dawkins R. 1976. The Selfish Gene. W. W. Norton and Company.

Fisher RA. 1941. Average excess and average effect of a gene substitution. Annals of Human Genetics 11: 53-63.

Frank SA. 1998. Foundations of Social Evolution. Princeton University Press.

Frank SA. 2013. Natural selection. VII. History and interpretation of kin selection theory. Journal of Evolutionary Biology 26: 1151-1184.

Gardner A, Grafen A. 2009. Capturing the superorganism: A formal theory of group adaptation. Journal of Evolutionary Biology 22: 659-671.

Gardner A, West SA. 2010. Greenbeards. *Evolution* 64: 25-38.

Gardner A, West SA, Barton NH. 2007. The relation between multilocus population genetics and social evolution theory. American Naturalist 169: 207-226.

Gardner A, West SA, Wild G. 2011. The genetical theory of kin selection. Journal of Evolutionary Biology 24: 1020-1043.

Grafen A. 1985. A geometrical view of relatedness. Oxford Surveys in Evolutionary Biology, 2: 28-89.

Grafen A. 2006. Optimization of inclusive fitness. Journal of Theoretical Biology 238: 541-563.

Grafen A. 2014. The formal Darwinism project in outline. Biology and Philosophy 29: 155-74.

Hamilton WD. 1964. The genetical evolution of social behaviour. Journal of Theoretical Biology 7: 1-52.

Hamilton WD. 1975. Innate social aptitudes of man: An approach from evolutionary genetics. Pages 133-155 in Fox R, ed. Biosocial Anthropology. Wiley.

Hölldobler B, Wilson EO. 2009. The Superorganism: The Beauty, Elegance and Strangeness of Insect Societies. W. W. Norton and Company.

Kitcher P. 1989. Explanatory unification and the causal structure of the world. Pages 410-505 in Kitcher P, Salmon WC, eds. Scientific Explanation. University of Minnesota Press.

Lehmann L, Keller L. 2006. The evolution of cooperation and altruism—a general framework and a classification of models. Journal of Evolutionary Biology 19: 1365-1376.

Lehmann L, Keller L, West S, Roze D. 2007. Group selection and kin selection: Two concepts but one process. Proceedings of the National Academy of Sciences USA 104: 6736-6739.

Lehmann L, Rousset F. 2010. How life history and demography promote or inhibit the evolution of helping behaviours. Proceedings of the Royal Society of London B: Biological Sciences 365: 2599-2617.

Lehmann L, Rousset F. 2014a. Fitness, inclusive fitness and optimization. Biology and Philosophy 29: 181-195.

Lehmann L, Rousset F. 2014b. The genetical theory of social behaviour. Philosophical Transactions of the Royal Society of London B: Biological Sciences 369: 20130357.

Lee JJ, Chow CC. 2013. The causal meaning of Fisher's average effect. Genetics Research 95: 89-109.

Mc Ginty SÉ., Lehmann L, Brown SP, Rankin DJ. 2013. The interplay between relatedness and horizontal gene transfer drives the evolution of plasmid-carried public goods. Proceedings of the Royal Society of London B: Biological Sciences 280: 20130400.

Marshall JAR. 2011. Group selection and kin selection: Formally equivalent approaches. Trends in Ecology and Evolution 26: 325-332.

Maynard Smith J. 1964. Group selection and kin selection. Nature 201: 1145-1147.

Maynard Smith J. 1976. Group selection. Quarterly Review of Biology 51: 277-283

Maynard Smith J. 1983. Models of evolution. Proceedings of the Royal Society of London B: Biological Sciences 219: 315-325.

Maynard Smith J, Szathmáry E. 1995. The Major Transitions in Evolution. Oxford University Press.

Michod RE. 1982. The theory of kin selection. Annual Review of Ecology and Systematics 13: 23-55.

Nowak MA, Tarnita CE, Wilson EO. 2010. The evolution of eusociality. Nature 466: 1057-62.

Nowak MA, Tarnita CE, Wilson EO. 2011. Nowak et al. reply. Nature 471: E9-E10.

Okasha S. 2006. Evolution and the Levels of Selection. Oxford University Press.

Okasha S, Paternotte C. 2012. Group adaptation, formal Darwinism and contextual analysis. Journal of Evolutionary Biology 25: 1127-1139.

Okasha S, Weymark JA, Bossert W. 2014. Inclusive fitness maximization: An axiomatic approach. Journal of Theoretical Biology 350: 24-31.

Okasha S. forthcoming. The relation between kin and multi-level selection: An approach using causal graphs. British Journal for the Philosophy of Science.

Pearl J. 2009. Causality: Models, Reasoning and Inference. 2nd ed. Cambridge University Press.

Price GR. 1970. Selection and covariance. Nature 227: 520-1.

Price GR. 1972. Extension of covariance selection mathematics. Annals of Human Genetics 35: 485-90.

Queller DC. 1984. Kin selection and frequency dependence: A game-theoretic approach. Biological Journal of the Linnaean Society 23: 133-143.

Queller DC. 1992. A general model for kin selection. Evolution 46: 376-380.

Queller DC. 2011. Expanded social fitness and Hamilton's rule for kin, kith and kind. Proceedings of the National Academy of Sciences USA 108: 10792-10799.

Rousset F. 2004. Genetic Structure and Selection in Sub-Divided Populations. Princeton University Press.

Rousset F, Lion S. 2011. Much ado about nothing: Nowak et al.'s charge against inclusive fitness theory. Journal of Evolutionary Biology 24: 1386-1392.

Salmon WC. 1989. Four Decades of Scientific Explanation. University of Minnesota Press

Sober E, Wilson DS. 1998. Unto Others: The Evolution and Psychology of Unselfish Behaviour. Harvard University Press.

Spirtes P, Glymour C, Scheines R. 2000. Causation, Prediction and Search. 2nd ed. MIT Press.

Taylor PD. 1990. Allele frequency change in a class-structured population. American Naturalist 135: 95-106.

Taylor PD, Frank SA. 1996. How to make a kin selection model. Journal of Theoretical Biology 180: 27-37.

Taylor PD, Wild G, Gardner A. 2007. Direct fitness or inclusive fitness: How shall we model kin selection? Journal of Evolutionary Biology 20: 301-309.

Taylor C, Nowak MA. 2007. Transforming the dilemma. Evolution 61: 2281-2292.

Traulsen A. 2010. Mathematics of kin- and group-selection: Formally equivalent? Evolution 64:316-323.

van Veelen M. 2009. Group selection, kin selection, altruism and cooperation: When inclusive fitness is right and when it can be wrong. Journal of Theoretical Biology 259: 589-600.

van Veelen M, Garcia J, Sabelin MW, Egas M. 2012. Group selection and inclusive fitness are *not* equivalent; The Price equation vs. models and statistics. Journal of Theoretical Biology 299: 64-80.

West SA, Griffin AS, Gardner A. 2008. Social semantics: How useful has group selection been? Journal of Evolutionary Biology 21: 374-383.

West SA, Gardner A. 2013. Inclusive fitness and adaptation. Current Biology 23: R577-R584.

Wilson EO. 2012. The Social Conquest of Earth. W. W. Norton and Company.

Wilson EO, Hölldobler B. 2005. Eusociality: origins and consequences. Proceedings of the National Academy of Sciences USA 102: 13367-13371.

Wilson EO, Nowak MA. 2014. Natural selection drives the evolution of ant life cycles. Proceedings of the National Academy of Sciences USA 111: 12585-12590.

Figures and boxes

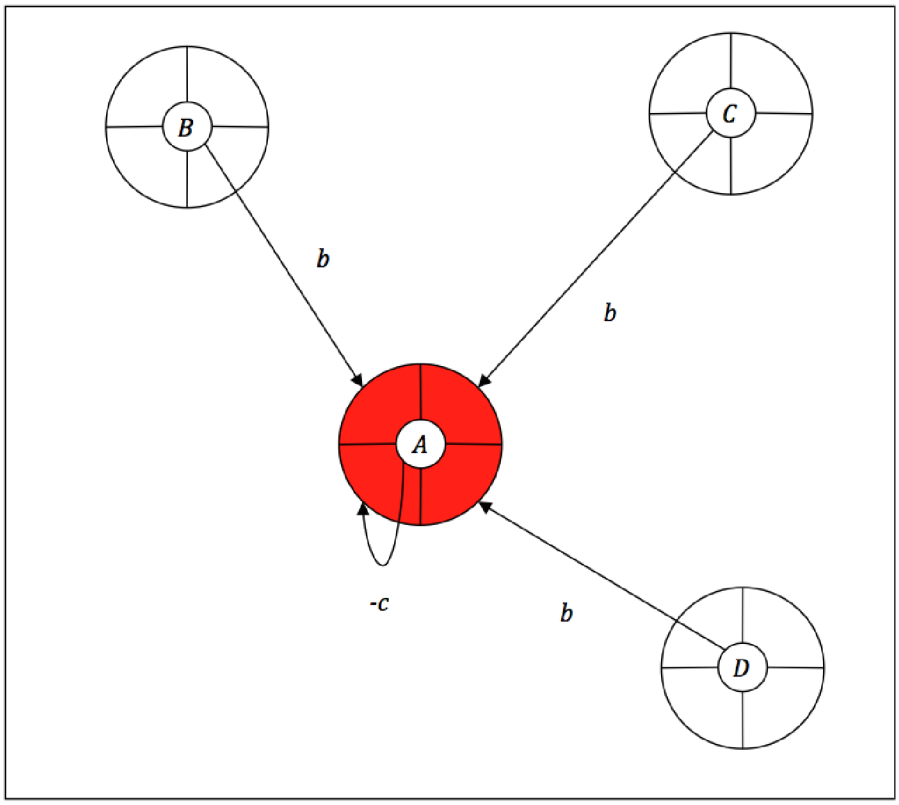


Figure 1: **Neighbour-modulated fitness.** In a neighbour-modulated fitness analysis, we ascribe to A those fitness components that correspond to its personal reproductive success. Some of these components are influenced by the behaviour of B, C and D (as shown by the arrows). A's total neighbour modulated fitness is an simple sum of these components ($3b$), plus a component corresponding to A's own influence on its reproductive success ($-c$), plus a baseline component independent of the character of interest.

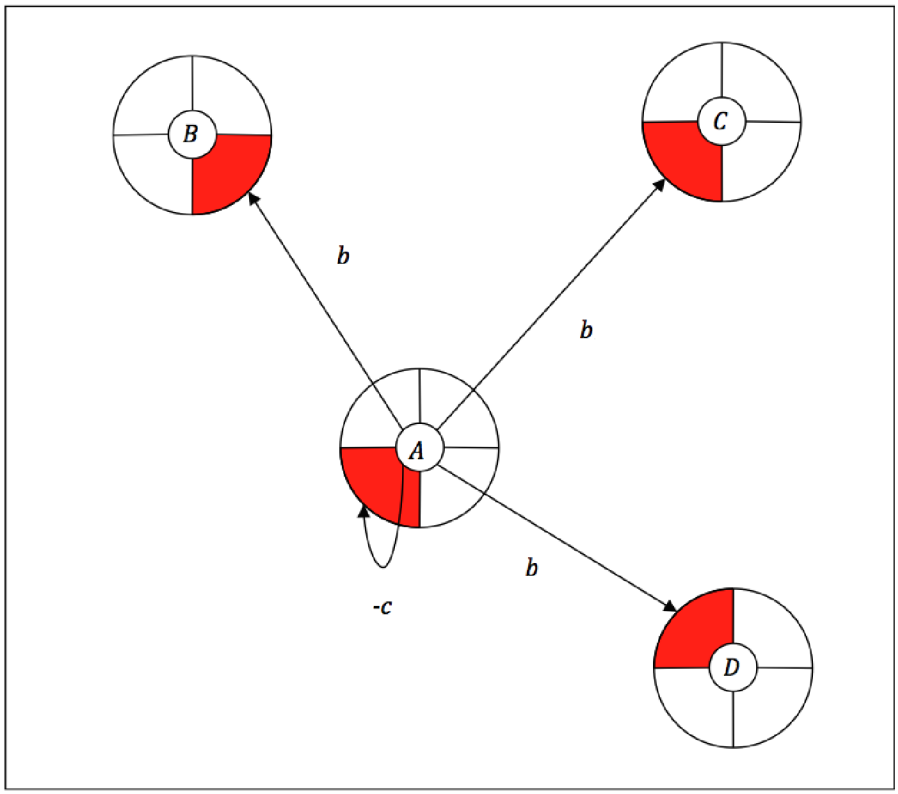


Figure 2: **Inclusive fitness.** In an inclusive fitness analysis, fitness effects are assigned to the actors whose behaviour was causally responsible for them. A therefore retains the effect $-c$ for which it responsible, but loses the $3b$ units of personal fitness it received by virtue of its interactions with B, C, and D. In compensation, it gains $3b$ units taken from the reproductive output of B, C and D. To calculate A's inclusive fitness, these new slices are weighted by the actor's relatedness to the recipient.

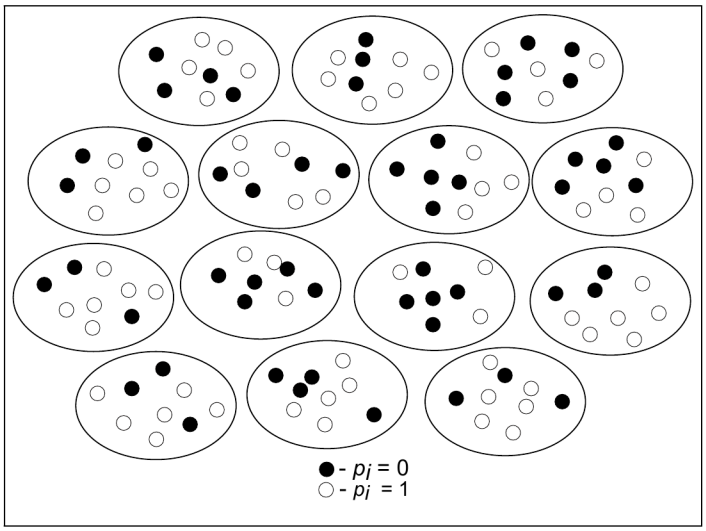
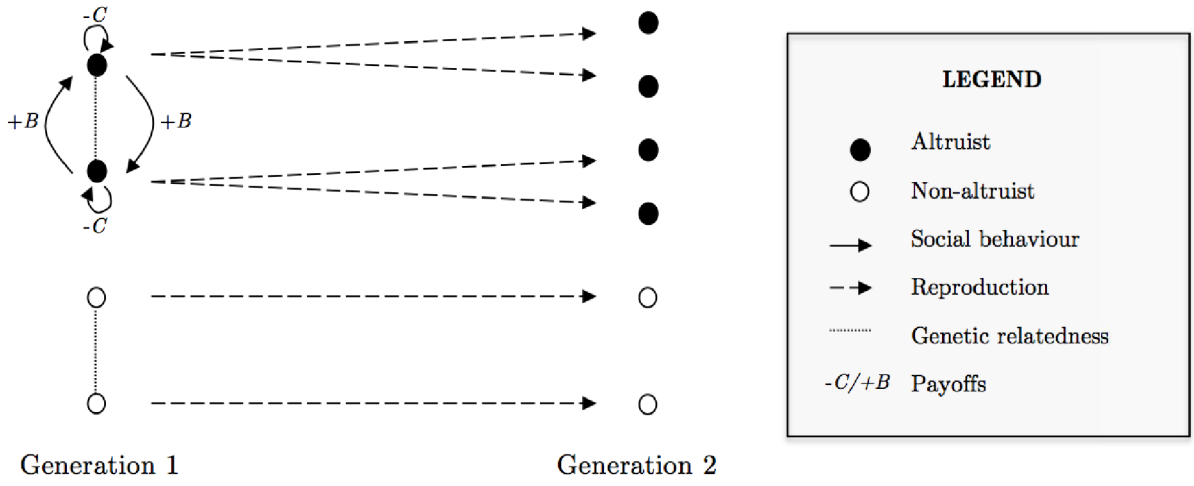


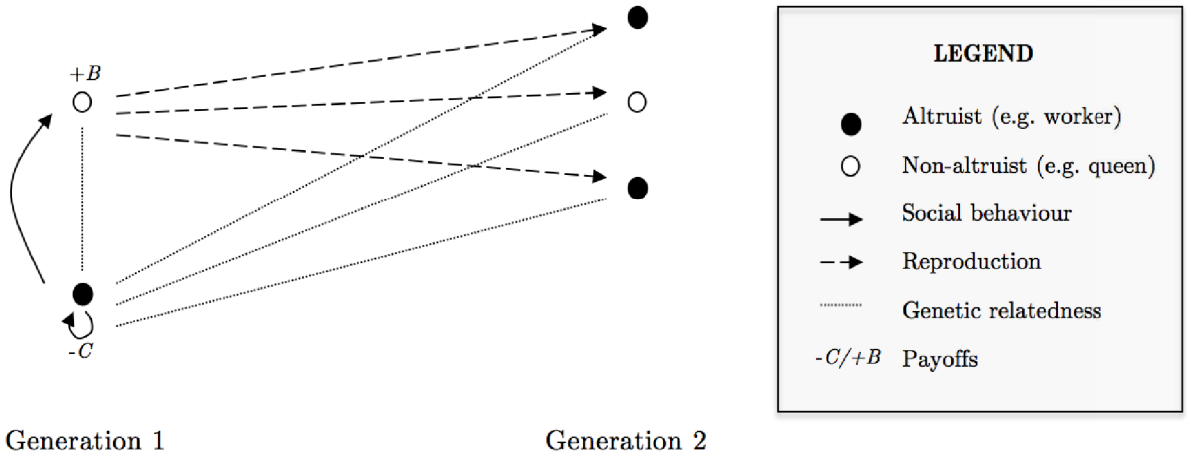
Figure 3: **Individuals in a group-structured population.** Black dots represent individuals with the allele of interest, white dots represent non-bearers, and the larger circles denote social groups.

Box 1: Two ways to conceptualize the role of relatedness



Picture 1: Relatedness leads to correlated interaction. Two altruists (black) confer a fitness benefit (B) on each other at a cost (C) to themselves. As a result, they are fitter overall than two nearby non-altruists (white). Genetic relatedness can give rise to such patterns of correlated interaction in a population, making altruists fitter (on average) than non-altruists.

650



Picture 2: Relatedness leads to indirect reproduction. An altruist (black) confers a fitness benefit (B) on a related recipient (white) at a cost (C) to itself. The recipient does not express the altruistic phenotype. However, it possesses conditionally expressed genes for altruism, which it transmits to some of its offspring (as indicated by the dotted lines, which show the genetic similarity between the actor and the recipient's offspring). The recipient thereby provides the actor with a means of 'indirect reproduction'—that is, an indirect route to genetic representation in the next generation.

Box 2: Kin selection approach

w_i = fitness of individual i

p_i = genetic value of individual i

p'_i = average genetic value of individual i's social partners

Write w_i as a multiple regression on p_i and p'_i:

$$w_i = \alpha + \beta_{wp.p'} p_i + \beta_{wp'.p} p'_i + e_i \tag{2}$$

Substitute equation (2) into (1) to yield:

$$\overline{w}\Delta\overline{p} = \big(\beta_{wp.p'} + \beta_{wp'.p}\beta_{p'p}\big)\mathrm{Var}(p) \tag{3}$$

where $\beta_{p'p}$ is the linear regression of p' on p.

Re-label $\beta_{wp.p'}$ and $\beta_{wp'.p}$ as '$-c$' and 'b' respectively, and $\beta_{p'p}$ as 'r', to give:

$$\overline{w}\Delta\overline{p} \;=\; \overbrace{(-c)\mathrm{Var}(p)}^{\text{direct effect}} \;+\; \overbrace{rb\mathrm{Var}(p)}^{\text{indirect effect}} \tag{4}$$

Equation (4) yields the generalized Hamilton's rule (HRG):

$$\Delta\overline{p} > 0 \text{ if and only if } rb > c \text{ (provided } \mathrm{Var}(p) \neq 0)$$

Box 3: Multi-level selection approach

p_{jk} = genetic value of j^{th} individual in k^{th} group

w_{jk} = fitness of j^{th} individual in k^{th} group

P_k = average genetic value of k^{th} group

W_k = average fitness of k^{th} group

The overall covariance between w and p, in the global population, can be written:

$$\text{Cov}(w_i, p_i) = \overbrace{\text{Cov}(W_k, P_k)}^{\text{between-group}} + \overbrace{E_k[\text{Cov}(w_{jk}, p_{jk})]}^{\text{within-group}} \quad (5)$$

652

where $\text{Cov}(W_k, P_k)$ is the covariance between the group means and

$\text{E}_k[\text{Cov}(w_{jk}, p_{jk})]$ is the average of the within-group covariances between w and p.

Substituting equation (5) into equation (1) yields:

$$\overline{w}\Delta\overline{p} = \overbrace{\text{Cov}(W_k, P_k)}^{\text{between-group}} + \overbrace{E_k[\text{Cov}(w_{jk}, p_{jk})]}^{\text{within-group}} \quad (6)$$

Equation (6) tells us that

$$\Delta\overline{p} > 0 \text{ if and only if } \text{Cov}(W_k, P_k) > -\text{E}_k[\text{Cov}(w_{jk}, p_{jk})]$$